A Video Indexing Approach Based on Audio Classification

A. Bugatti, R. Leonardi and L. A. Rossi
Department of Electronics for Automation, University of Brescia
Via Branze 38, I-25123 Brescia – Italy
{alex@imago., leon@, lrossi@}ing.unibs.it

Abstract

This paper presents a video indexing approach based only on audio classification. Indeed, we apply to an audio-visual document a set of methods for partitioning the associated audio data into homogeneous segments. The aim is to highlight semantically relevant items of a multimedia document by relying only on simple audio processing techniques. A simple algorithm to identify audio segments belonging to silence, music, speech and noise classes has been proposed.

1. Introduction

Nowadays most multimedia material is provided in (compressed) digital format. A still missing element is a standard procedure to characterize the content of the material, which may constitute an index to the infomation. This is essential to provide the framework for an effective navigation through the multimedia repository, and the retrieval of relevant information. In this context, the International Standard Organization (ISO) started in October 1996 a standardization process for the description of the content of multimedia documents, namely MPEG-7: the "Multimedia Content Description Interface" [2], [3].

The issue of allowing for possible automatic procedures to semantically index audio-video material represents a very important challenge. Such procedures create indices of the audio-visual material, some of which should have the capability to characterize the temporal structure of the multimedia document, possibly even at a semantic level. In this work, we see the first stage of such decomposition as a series of consecutive segments, which are coherent from a certain point of view, initially low level. By organizing the degree of coherence, according to higher level criterion, it is possible subsequently to construct a hierarchical representation of information (in terms of semantics), so as to create a Table of Content description of the document. Such a representation is indeed quite suited for navigation purposes, thanks to the multi-layered summary that can be generated.

Traditionally, the most common approach to create an index of a video document has been based on the automatic detection of camera record changes with their associated editing effects [1]. This kind of approach has generally demonstrated satisfactory performance and lead to a good low-level temporal characterization of the visual content. However the reached semantic level remains poor since the description is very fragmented considering the high number of shot transitions occurring in typical audio-visual programs.

Alternatively, there have been recent research efforts to base the analysis of MM documents by a joint audio and video processing so as to provide for a higher level organization of information. In [9], [10], these two sources of information have been jointly considered for the identification of simple scenes that compose a multimedia program. The video analysis associated to cross-modal procedures can be very computationally intensive (by relying, for example, on identifying correlation between non-consecutive shots). Moreover, it has been limited up to now to ad hoc procedures for identifying very peculiar types of scenes, such as dialogues, with average performance. On the other hand, audio information carries out by itself a rich amount of semantic significance [13]. Hence, in this work, we propose a set of methods for multimedia indexing which makes use only of audio analysis. The basic assumption behind this approach is that a relevant segment of multimedia document from an audio semantic point of view is meaningful also according to the semantics of the associated video [13]. Our aim is to have a set of low cost but effective methods for audio-video indexing.

The contribution is organized as follows. Section 2 presents the proposed algorithm, while Section 3 shows some simulation results. Finally conclusions are drawn in Section 4.

2. The audio classification algorithm

According to several researches in the field of audio segmentation [5], [7], [9], [13], a reasonable audio model is composed of four classes: silence, music, speech and noise. Our approach is based on a

This work has been partially founded by the European ESPRIT project 28798 AVIR (Audio-Visual Indexing and retrieval for non IT expert users).

hierarchical decomposition: at each level a different algorithm is applied to carry out a finer segmentation of the audio stream (see Figure 1).

In the first level, the audio file is separated in silence and non-silence segments. In a second phase, the non-silence segments are separated into the remaining three other classes.

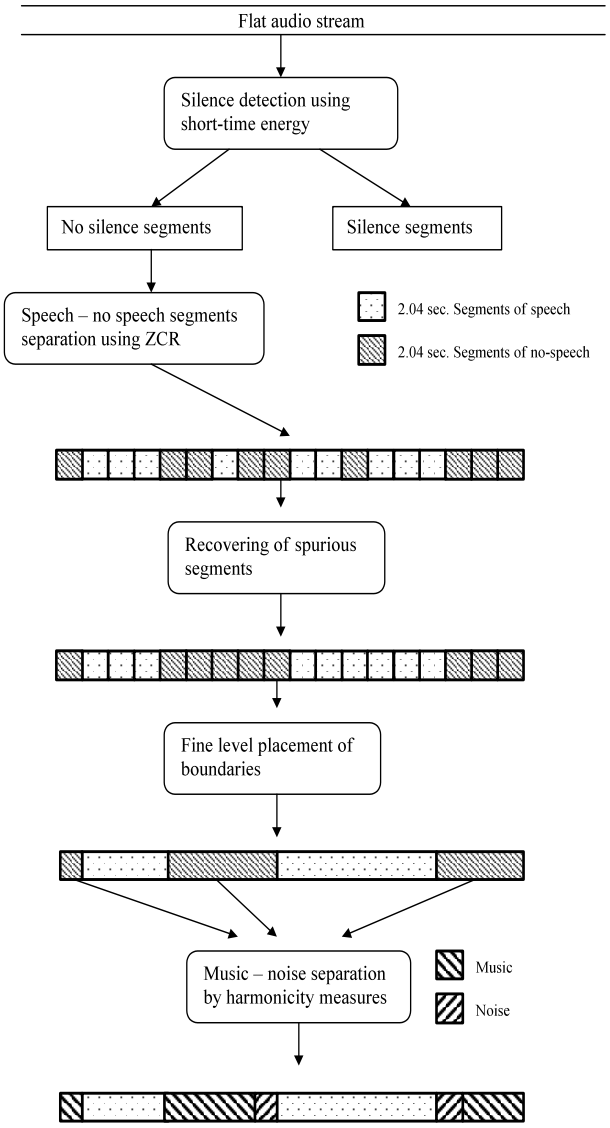

Figure 1: Audio hierarchical decomposition.

Our silence detection procedure uses a statistical pattern classification algorithm. By representing the silence/non-silence classes as vectors in $\Re_n$, the aim is to select adequate features and represent them as the vector components. If to each class the feature vectors define distinctive clusters, the classification problem is simple, as a candidate vector belongs to a given cluster, thus to the associated class. The short-time energy function provides an adequate measure of the amplitude variations of the signal and it is used in a few silence detection frameworks [4], [6]. In this case, the feature vectors, $\mathbf{x}$, are purely one-dimensional. The classifier divides the space into only two regions R_1 and R_2. When the pattern $\mathbf{x}$ falls into the region R_1, $\mathbf{x}$ is classified as belonging to w_1 (i.e. the silence class), otherwise $\mathbf{x}$ is classified as belonging to w_2 (i.e. the non-silence class). The developed algorithm, presented hereafter does not require any a priori information on the silence/non-silence distribution. It performs, whenever it is possible, an initial training in order to evaluate the statistics of the local energy level of the background noise, which defines the initial condition for silence. The training takes place on the first few samples of audio signal (typically 0.4 sec.). Obviously these samples have to be representative of the background noise for the entire audio stream: if the first few seconds of the audio stream are not in silence, an estimate of the background noise is obtained by a random selection of frames with low energy level. Following anyone of these two estimation strategies, the statistics of the silence energy level distribution can be dynamically updated once the processing of the audio stream takes place. This avoids future misclassification due to changes in statistics. The only necessary assumption is that the background noise does not exhibit abrupt changes in statistics. If the background noise can be considered wide sense stationary, at least during short time intervals, a dynamical update can be reasonably achieved.

As indicated previously, once silence and non-silence segments have been separated, the non-silence segments are further processed to identify speech, music and noise sub-segments. For such a purpose ,we rely on speech characteristics to discriminate it from music and noise. Speech shows a very regular structure while music and noise do not present such a regularity. Indeed the speech is composed by a succession of vowels and consonants: while the vowels are high energy events with most of their spectral energy distributed around low frequencies, the consonant are noise-like with their spectral energy distributed mainly towards the higher frequencies. Saunders [12] uses the zero-crossing rate (ZCR) to identify this behavior. An example of such behavior is shown in Figure 2.

The audio file (without the silence segments) is partitioned into segments of 2.04 seconds, so that each of them is composed of 150 consecutive non overlapping frames. These values have been selected so

as to have a meaningful number of samples in each frame for a 22050 Hz sample frequency. Thus, each frame contains 300 samples, and by selecting 150 frame interval for statistical estimates, we obtain an adequate trade-off between the quasi-stationary properties of the signal and a sufficient length to evaluate the ZCR. For every frame, the value of the zero-crossing rate is calculated using the definition in [8]. This window of 150 values of the ZCR is used to estimate the following parameters:

- *ZCR Variance*: which indicates the dispersion with respect to the mean value.
- *ZCR Third order moment*: which indicates the degree of skewness with respect to the mean value.
- Difference between the number of ZCR samples which are above and below the mean.

Thus with each segment of 2.04 seconds, a 3-dimensional vector can be constructed. To achieve the separation between speech and the other classes using a computationally efficient implementation, a multivariate Gaussian classifier has been used.

At the end of this step we have a set of consecutive segments labeled as speech or non-speech. To clean up the results of the Gaussian classifier, a post-processing is applied. It is justified to an empirical observation: the probability to observe a single segment of speech surrounded by music or noise segments (i.e. non-speech segments) is very low and vice versa. Therefore a simple processing, such as a median filter, is applied to change the labels of these spurious segments.

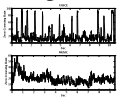

Figure 2: The ZCR evolution for voice and music segments.

The resolution of the different segment boundaries is fixed by the 2,04 sec window duration, which is inherently on the nature of the ZCR algorithm. Obviously these boundaries are not very accurate, thus there is the need of a fine level analysis of the segments across the boundaries, so as to determine a more accurate positioning. The ZCR values of the neighboring segments are processed to identify the exact transition between two consecutive speech and non-speech segments. A new measure is computed from their associated ZCR values as follows:

$$y[n] = \frac{1}{P}\sum_{m=n-P/2}^{n+P/2}(x[m]-\bar{x}_n)^2 \quad \text{with} \quad P/2 < n < 600 - P/2 \ ,$$

where x[n] is the n-th ZCR values of the segments and $\bar{x}_n$ is defined below:

$$\bar{x}_n = \frac{1}{P}\sum_{m=n-p/2}^{n+P/2} x[m] \ .$$

P defines the duration of a short window, typically set to 30. Thus, y[n] is an estimate of the ZCR variance over such short windows of length P. A low pass filter is applied to this variance estimate to obtain a smoother measure and finally a peak extractor is used to precisely locate the transition between speech and non-speech.

The last step is the music-noise separation in non-speech segments. Generally, the music can be considered composed of harmonics, where a harmonic sound consists of a series of major frequencies components including the fundamental frequency and those which are integer multiples of it. According to this observation we decide to obtain a measure of the degree of harmonicity so as to discriminate between music and noise (as most environmental sounds are non-harmonic like). A simple pitch detector based on the FFT applied over a 1024 sample window is used to achieve this objective.

3. Simulation results

We have developed a software interface (Figure 3), which allows to visualize the classification results so as to ease any comparison between the automatic classification procedure and a subjective classification, obtained by listening to the audio stream.

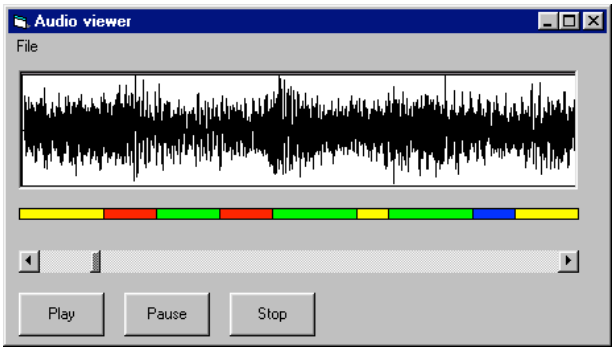

Figure 3: The interface used to test the classification results.

At present, we have carried only preliminary simulations, but the so far obtained results have shown good performance in classifying audio and consequently have provided meaningful temporal characterization of audio-visual segments. This has been verified subjectively by identifying the semantic relevance of the obtained classification for various types of programs, such as movies, documentaries, broadcast programs,…

4. Conclusion

We have developed an alternatively approach to the multimedia material segmentation based only on audio analysis. Our methods have a low complexity and they are very attractive from a computational point of view.

The first results show that this approach carries out a meaningful segmentation, even with respect to the semantics of the associated video. The results also exhibit some classification errors in presence of high background noise or other particular situations. These will be addressed in the future by adding peculiar feature parameters to avoid the results of the misclassification. More attention is needed also to further investigate the semantic significance of this lw level audio processing in terms of the segmentation of the complete audio-visual document.

References

[1] N. Adami and R. Leonardi. Identification of editing effects in image sequences by statistical modelling. In *Proc. Picture Coding Symposium '99*, Portland, OR, U.S.A., Apr. 1999.

[2] MPEG Requirement Group. MPEG-7: Context and objective. *ISO/IEC JTC1/SC29/WG11* **N2460**, MPEG98, Atlantic City, USA, Oct. 1998.

[3] MPEG Requirement Group. MPEG-7: Requirements. *ISO/IEC JTC1/SC29/WG11* **N2461**, MPEG98, Atlantic City, USA, Oct. 1998.

[4] De Souza, P. A Statistical Approach to the Design of an Adaptive Self-Normalizing Silence Detector. *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-31, 3 (Jun. 1983), 678-684.

[5] J. Foote. "A similarity measure for automatic audio classification". *In Proc. AAAI'97 Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, 1997.

[6] Lamel, L.F., Rabiner, L.R., Rosenberg, A.E., and Wilpon, J.G. An Improved Endpoint Detector for Isolated Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-29, 4 (Aug. 1981), 777-785.

[7] C. Montaciè, M. Caraty, A Silence/Noise/Music/ Speech Algorithm, *International Conference on Spoken Language Processing*, Sidney 1998.

[8] L. Rabiner, R. Schafer, Digital Processing of Speech Signals, Prentice Hall, Alan Oppenheim editor

[9] C. Saraceno. Content-based representation and analysis of video sequences by joint audio and visual characterization. PhD thesis, Brescia, 1998.

[10] C. Saraceno and R. Leonardi: Indexing audio-visual databases through a joint audio and video processing. *International Journal of Imaging Systems and Technology*, 9(5):320-331, Oct. 1998.

[11] C. Saraceno and R. Leonardi. Identification of story units in audio-visual sequences by joint audio and video processing. In *Proc. International Conference on Image Processing 1998*, Chicago, IL, U.S.A., Oct. 1998.

[12] J.Saunders, Real Time discrimination of broadcast music/speech. *In Proc. 1996 ICASSP*, pages 993-996, 1996.

[13] T. Zhang and C.-C. Jay Kuo, " Audio-Guided Audiovisual Data Segmentation and Indexing", *IS&T/SPIE's Symposium on Electronic Imaging Science & Technology -- Conference on Storage and Retrieval for Image and Video Databases VII*, SPIE Vol.3656, p316-327, San Jose, Jan. 1999.